II. Interactions with the SUMEX-AIM Resource

II.A Use of programs via SUMEX

CONGEN Workshops

In early December, 1978, we held at Stanford a series of mini-workshops on the use of an exportable version of the CONGEN program. Invitees included members of the chemical and biochemical community who are actively engaged in solving the structures of unknown chemical compounds encountered in research in industrial, academic and government research laboratories. The primary purpose of these workshops was to introduce experts in the field of structure elucidation to the first version of the exportable program. These persons were chosen for their chemical and biochemical expertise; few had significant experience with computers previously. Thus, they represented what we think is a good cross-section of the community of potential users of CONGEN. We held three three-day sessions of the workshop so that we could offer access to a computer terminal for all the persons at one session and so that we could provide close supervision and assistance as they began to learn and use CONGEN. We also implemented a recording scheme so that an interactive session at the terminal could be recorded as a text file and available after the problem was completed for close scrutiny for the chemist and for ourselves. Such scrutiny reveals, for example, common difficulties in certain portions of the user interaction thereby pointing out areas for improving the interaction.

To help alleviate the impact on SUMEX, the CONGEN program was brought up on the Rutgers-AIM (RUAIM) system and half the participants used it there via TYMNET. There were three users (average) on SUMEX and three on RUAIM during our terminal sessions. Nevertheless, our impact on SUMEX was very large and control on other users had to be applied during our terminal sessions.

Although the version of CONGEN used in the workshops was not complete, enough of the program existed in close to final form to allow us to fulfill our other purposes. We wanted to ensure that any remaining program errors could be detected and fixed prior to making the program more widely available. The best way we have found to do this once a program is essentially debugged is to confront the program with a wide variety of problems from many different users. We also wanted to determine if there were major deficiencies in any part of the program which made it difficult to understand or use. Eliminating such deficiencies would ensure that an exported version would meet the needs of the persons attending the workshop, i.e., that some minimum standards of acceptability could be determined and met. Finally, we needed to determine the computing facilities available to this group and in detailed discussions to explore opportunities for export to their own laboratories. This allows us to set some priorities on developing versions for various makes of computers.

Conclusions from the Workshop

There are several conclusions which can be drawn from the workshop experience. The reaction of all persons attending the workshop was very positive, not only concerning organization and intellectual stimulation, but also with the problem-solving capabilities of the program. The following are major positive aspects of the workshop experience:

a) we were able to meet our goal of demonstration of exportability by utilizing CONGEN on two different computers during the workshop;

b) every participant found the program of sufficient utility to express an interest in obtaining a version in some way for his or her own laboratory;

c) the interface to CONGEN, extensively modified based on experience with the old version of the program, proved much simpler to use, much more chemically logical and consistent and much more helpful to the user in providing guidance and error checking;

d) several new problems were analyzed successfully at the workshops, either by verification of the unambiguous nature of the structural assignment or by obtaining a list of candidate solutions to guide further experimentation;

e) installation of the exportable version has been completed successfully at two different sites, Lilly Research and Smith, Kline and French Research, and several more will follow in the next two months.

There are some common criticisms expressed by the persons attending the workshop which, in our opinion, represent points of focus for the remainder of the grant period and for a renewal application. Briefly, the major deficiencies were as follows:

a) The requirement of specifying non-overlapping structural units is non-intuitive and thus unnatural. Other programs, like CONGEN, share this difficulty, but we are in a position to remedy it based on recent research so that future versions may be easier to use;

b) The program is very complex and lacks sufficient documentation or internal 'help' facilities. We recognize this and to some extent it is a reflection of the lack of maturity of the new version. We plan to provide better on-line help facilities accessible from within the program and a much more comprehensive program guide with examples.

c) The teletype oriented drawing program produces some drawings which are difficult (if not impossible) to interpret. Providing the chemist with a connection table of such drawings, as we can do currently, is no long-term solution. Here we face the problem of diminishing the exportability of the program if we restrict its use to certain types of graphics terminals (there are many types, each requiring different programs to operate). Currently there is no graphics terminal which is competitive in price to character-oriented terminals. One way to solve this problem is to encourage collaborators to provide their own graphics packages which we can then in turn offer to others.

CONGEN Workshop attendees, affiliation, and interests

1) Dr. Henry Stoklosa, E.I. DuPont de Nemours. Dr. Stoklosa has been affiliated with a group at DuPont involved with computer applications to chemical problems, including computer-aided organic synthesis.

2) Dr. G.W.A. Milne, National Institutes of Health. Dr. Milne is currently in charge of the National Institutes of Health contribution to the NIH/EPA Chemical Information System. His interests included not only evaluation of the utility of the program but also exploration of ways in which CONGEN might be interfaced to the Chemical Information System.

3) Dr. William Brugger, International Flavors and Fragrances. Dr. Brugger represents the key person at IFF Research responsible for computer applications in their laboratories. Structure elucidation is a major activity of this company not only in analysis of natural and synthetic products but also in assessing the relationships between chemical structure and toxic properties affecting human health.

4) Dr. Douglas Dorman is head of the NMR laboratory at Lilly Research Laboratories and works closely with mass spectroscopists and other chemists in solving structures of a variety of compounds related to existing or new products. Dr. Dorman has been familiar with the "old" (non-exportable) version of CONGEN and thus was able to critique the new program not only on its merits but also on comparison with the old version.

5) Dr. Jon Clardy, Cornell University. Dr. Clardy is a recognized leader in development and applications of the technique of X-ray crystallography in structure elucidation. His attendance of the workshop was based on an interest in learning about alternative, computer-based approaches to the problem. As his letter points out, the ability to use CONGEN to help solve structures before expenditure of time and effort in X-ray analysis would be an important benefit.

6) Mr. In Ki Mun, Cornell University. Mr. Mun attended the workshop representing Prof. Fred McLafferty at Cornell. Prof. McLafferty's group has had for many years an interest in use of computer techniques to help solve structures, based primarily on mass spectral data. His research in this area has led to programs which suggest the presence of functionalities in an unknown molecule. CONGEN can, in principle, complete such a schema for analysis by piecing together the inferred functionalities.

7) Dr. Reimar Breuning, Munich. Dr. Breuning learned of the existence of the workshops from discussions with Prof. Djerassi at the IUPAC meeting on natural products. He is actively involved in natural products structure elucidation at Munich.

8) Dr. David Lynn, Columbia University. Dr. Lynn attended the workshop representing Prof. Koji Nakanishi, the latter a recognized expert in the area of structure elucidation of a number of classes of natural products of relevance to human health. Dr. Lynn is to act as the focal point for introduction of the computer methods to that research group.

9) Dr. Y. Gopichand, University of Oklahoma. Dr. Gopichand attended the workshop representing the marine natural products group of Prof. Francis Schmitz. This group specializes in structure elucidation of halogenated terpenoid molecules possessing a variety of biological activities and marine sterols representing intermediates or end products in steroid biosynthesis.

10) Ms. Wendy Harrison, University of Hawaii. Ms. Harrison attended the workshop representing the marine natural products group of Prof. Paul Scheuer at Hawaii. This group is engaged in structure elucidation problems similar to those encountered in Prof. Schmitz's laboratory, although focus is on different classes of organisms.

11) Dr. Laszlo Tokes and Dr. Michael Maddox, Syntex Research. Drs. Tokes and Maddox are, respectively, in charge of the mass spectrometry and NMR laboratories at Syntex Research. They are responsible for the majority of structure elucidation problems which rely on physical methods. Their interest in CONGEN is that it might help them solve certain problems in less time than required by manual methods.

12) Dr. John Figueras, Kodak Research Laboratory. Dr. Figueras attended representing the Analytical Sciences Division of Kodak's Research Laboratory. This division is responsible for data collection and analysis in support of the structure elucidation activities of the Laboratory including not only new developments in the photographic process but also the new technology of thin-film bound enzymes systems for clinical analyses.

13) Dr. Charles Snelling, University of Illinois. Dr. Snelling attended the workshop representing Prof. Kenneth Rinehart in the Chemistry Dept. at Illinois. Prof. Rinehart is also an acknowledged expert in structure elucidation with emphasis on macrolide antibiotics, halogenated terpenoids and other classes of natural and synthetic products of relevance to human health problems.

14) Dr. Gilles Moreau, Roussel UCLAF. Dr. Moreau attended the workshop representing the French pharmaceutical concern Roussel UCLAF. This company maintains an active group in computer applications in chemistry and wished to evaluate CONGEN for its use in their structural problems.

15) Prof. Andre Dreiding, Zurich. Dr. Dreiding has been interested in both the problem-solving and the pedagogical aspects of CONGEN for some time. He had previously used the old version and was gratified to see the improvements in the new version.

16) Dr. James Shoolery and Dr. Michael Gross, Varian Associates. Dr. Shoolery is in charge of Varian's NMR application laboratory and Dr. Gross is in charge of computer software for Varian's NMR/computer systems. Their respective interests match their responsibilities.

17) Dr. Daniel F. Chodosh, Smith, Kline and French. Dr. Chodosh was not actually invited to the workshop, but happened to visit our group during one of the sessions. He was sufficiently impressed that he procured a tape to carry away a copy of the program with him. He has now been supplied with a version of CONGEN for the PDP-10 and has it running at the SKF research laboratories in Philadelphia.

Recent GUEST access

In addition, the following persons during the past year have asked for information about and access to CONGEN. For the most part we have granted access

through the GUEST directory, setting up an account only for those users with more than occasional log-ins.

Dr. David Cowburn
Physical Biochemistry
The Rockefeller University
New York City

Prof. Cowburn now has an account on SUMEX in the EXODENDRAL account and is interested in extensions of CONGEN relevant to his work on peptide conformational analysis.

Douglas Henry
School of Pharmacy
Oregon State University
Corvallis, Oregon

He has been sent our programs for structure drawing for use on his own computer.

The following have asked for and received information on access to CONGEN at SUMEX via the GUEST facility.

Dr. H. Kating
Institut fur Pharmazeutische Biologie
Der Universitaet
Bonn, Germany

Dr. Adalbert Kerber
Lehrstuhl D fur Mathematik
Aachen, Germany

Dr. Brenda J. Kimble
Radiobiology Laboratory
University of California
Davis, California

Dr. J. Neubuser
Lehrstuhl D fur Mathematik
Aachen, Germany

Dr. George Padilla
Dept. of Physiology
Duke University Medical Center
Durham, N.C.

Dr. W. Sieber
Sandoz Ltd.
Basel, Switzerland

Dr. Babu Venkataraghavan [We also helped him bring up the Fortran draw
Lederle Laboratories program on the DEC-10 system at Lederle]
Pearl River, New York

Dr. Stephen Wilson
Dept. of Chemistry
Indiana University
Bloomington, Indiana

Prof. Glenn D. Prestwich
Dept. of Chemistry
State University of New York at Stony Brook
Stony Brook, New York 11794

Dr. Andrew Stuper
RCG Group
Rohm and Haas
Norristown and McKean Roads
Springhouse, Pa. 19477

Prof. E. F. Domino
Department of Pharmacology
M6322 Medical Sciences Building
Ann Arbor, Michigan 48109

Prof. John D. Roberts
Div. of Chem. and Chem. Engr.
Calif. Inst. of Tech.
Pasadena, Calif. 91125

In addition there has been further use of CONGEN by several of the workshop participants who are only able to access the program by remote connection to SUMEX.

II.B. Interaction with other SUMEX-AIM projects

We continue to have interaction with other SUMEX projects as well as other resources funded by NIH. For example, we have continual contact with the SECS project headed by W. T. Wipke at Santa Cruz. Prof. Wipke and several of his co-workers spend time with us here at Stanford. Dr. Martin Huber of this group was an informal participant at one of our Workshops held last December. Drs. Smith and Wipke have collaborated in organizing a symposium on computer handling of structural information at the American Chemical Society meeting in Hawaii in April, 1979. The proximity of our groups, both geographically and especially through SUMEX, encourage this sharing of ideas on related areas and we expect the cooperation to continue as long as SUMEX exists to facilitate it.

We also interact with the MOLGEN project at Stanford, primarily in the area of experiment planning as it relates to structure determination. Although the structures which MOLGEN is trying to determine are in many ways more complex, requiring different representations in the computer, our methods for structure generation have provided some insights for them and we have been guided in our initial attempts toward experiment planning by their much more intensive effort.

We have interacted with the Rutgers-AIM resource during our workshops in order to distribute the heavy computing load and to locate our newly exportable CONGEN program on another machine. As indicated in the previous section, this

experiment was completely successful and demonstrates one way in which the two resources can work together.

II.C. Critique of Resource Management

Our major problem continues to be one of insufficient SUMEX capacity for production use of our applications programs, particularly during prime time. This reflects itself in our inability to guarantee high-quality, interactive access to our collaborators, who have structures of important biomolecules to solve and who can use the current version of CONGEN or a related program to help solve them. It is also reflected in our own inability to carry out detailed (and frequently time-consuming) tests of experimental versions of new programs within our own group. Thus, not only do applications suffer, but new developments as well. The gradual increase in difficulty in answering the demands of both development and applications, in a period of steadily increasing load on the existing SUMEX facility, has prompted us to propose a dedicated machine for our own project, described in more detail in Section III, closely interdigitated with the SUMEX resource. In this way we hope to divert all applications use of our programs to the dedicated machine, thus decreasing our load on SUMEX.

Other than this issue, we have no complaints. The system and its support staff continue to be extremely cooperative in aiding us in our collaborative efforts, in providing assistance on problems related to languages, editors and other system-supported facilities and in advising us when we consider new directions in our own work.

We do have one request. There is currently no mechanism for keeping statistics on GUEST use of our programs. Because we have been asking most of our collaborators to access SUMEX via the GUEST directory, it is important for us to know who has been logged on, what programs they have been running and how much time they have consumed. Although we do have a recording mechanism to keep track of some use, a general capability for monitoring use of GUEST would be very helpful to us.

III. RESEARCH PLANS (8/79 - 7/81)

III.A. Long Range Goals

We will continue to develop our programs for computer-assisted structure elucidation and pursue the dissemination of these programs to the biomedical community. Specifically, we will further develop the ability of our programs to take account of the three-dimensional nature of chemical structures. Nearly all properties of compounds relevant to biomedicine require this information. We will further develop our GENOA program for interpretation of constraints which may involve overlapping substructures. This will represent a significant advance to the intelligence of our programs by allowing the program to take information from the chemist in a much more intuitive manner. We will also further develop programs to aid in spectral interpretation and prediction and experiment planning.

Our current grant period ends on April 30, 1980. We have submitted a renewal proposal for the June 1 deadline, for funding for a five-year period

5/1/1980 - 4/30/1985. An important aspect of this renewal proposal is our relationship to SUMEX. This new proposal is again for Resource Related Research, but now related to the SUMEX resource rather than the mass spectrometry laboratory as it is now. We have proposed purchase of a dedicated computer system as part of the proposal, to be interfaced with SUMEX in order to shift production use of our applications program to the new machine. Our relationship with SUMEX involves this computer as an experiment in resource sharing for programs which have matured to the point where outside collaborators can make fruitful use of the programs in their own biomedical research.

III.B. Justification and Requirements for Continued SUMEX Use

Our continued funding from NIH attests to the biomedical relevance of our effort and the importance of our research. We foresee an increased relevance to biomedicine because of our further emphasis on three-dimensional chemical structures and increased efforts at dissemination of working programs to the biomedical community. Important aspects of our research involve ongoing and planned, future collaborative efforts involving applications of our newly developed techniques to structure/property relationships of several classes of biologically important molecules, including peptides, sterols and other marine natural products, and opiate agonists and antagonists. SUMEX is our means of carrying out this research.

The SUMEX resource provides an excellent and reliable "critical mass" of services, software, and programs for further development of our programs. SUMEX has been, and will continue to be, our only source of source of computational support for our own work and that of our collaborators, at least until action is taken on our renewal proposal. Even should our renewal be funded in total, we plan on continued utilization of SUMEX for our development efforts. The system we propose to purchase is insufficient to meet all our computational needs and is heavily dependent on SUMEX to provide needed peripheral equipment, access to a variety of languages, text editors, etc., and a support staff to aid us in our development efforts when SUMEX-related problems arise.

III.C. Other Computational Resources

As mentioned in previous sections, we have, in a renewal proposal submitted to he NIH, requested additional computational resources as part of our new relationship to the SUMEX resource. The proposal is to purchase a stripped-down (in terms of peripheral equipment) Digital Equipment Corp. VAX-11/780 computer for the specific purpose of providing a highly-responsive, interactive environment for production use of our applications programs such as CONGEN and GENOA. This will ease significantly our demands on the already overburdened SUMEX machine. Our plans, however, are dependent on SUMEX for continued development, so that both computational resources are important for the conduct of the research as proposed.

III.D. Recommendations for the Future

Now that we have developed a plan for resource sharing involving purchase of hardware for our own group as an augmentation to the SUMEX facility, we are hoping that this effort alone will help ease the load average on the SUMEX machines. If there were some way to facilitate the same approach for other

research projects using SUMEX, given that they have matured to the point where outside collaborators desire access, then the entire community would benefit. We are hoping that, if our efforts are successful, we will act as a model to guide others in the same direction. A great deal of development work can be done on SUMEX even at high load averages. It is when the programs which result from this development begin to be applied to real-world biomedical problems that SUMEX response begins to deteriorate, and then everyone suffers from an overloaded machine.

4.2.4 HYDROID Project

HYDROID - Studies in Distributed Processing and Problem Solving

Prof. Gio Wiederhold
Depts. of Computer Science and Electrical Engineering
Stanford University

The potential of multi-processor networks is nearly universally appreciated and many research groups are working either on theoretical related issues or on actual implementations. Several medically related groups are expending effort on multi-processor developments (e.g., U. of Wisc., U. of Texas at Austin, etc), but we believe that a number of basic system design issues should be resolved first. These groups have aggregate resources much greater than we currently have. The demands of the required simulations are such that usage of SUMEX is not feasible. As noted in the previous year's report we have in fact used SUMEX mainly as a means of communication and not as a computational resource. One outcome of our demonstrated interest is the initiation of a joint project with IBM on distributed computing at Stanford, although here we are talking about a dozen, rather than hundreds or thousands of processors. A proposal to develop intelligently communicating databases for chronic disease management will use these facilities, as well as DEC equipment at Stanford's CS dept. and microprocessor based systems.

Two Ph.D. theses were produced under the aegis of this project. Reid Smith's "Contract Net" provides a powerful framework for the study of interacting parallel processes. Hector Garcia-Molina's work on the "Performance of Update Algorithms in a Distributed Database" has lead to efficient algorithms for the maintenance of consistency in the databases in multiprocesses or network.

1. Garcia-Molina, Hector: "Performance Comparison of Update Algorithms for Distributed Databases," Technical Note 143, CSL, Stanford University, Sept. 1978.

2. Garcia-Molina, Hector: "Performance Comparison of Update Algorithms for Distributed Databases, Part II" ; CSL Tech Note 146, Dec. 1978.

3. H. Garcia-Molina, "Crash Recovery in the Centralized Locking Algorithm"; November 1978, to appear as a Tech Note.

4. H. Garcia-Molina, "Restricted Update Transactions and Read Only Transactions"; January 1979, to appear as a CSL Tech Note.

5. H. Garcia-Molina, "Partitioned Data, Multiple Controllers, and Transactions with an Initially Unspecified Base Set"; February 1979, to appear as a CSL Tech Note.

6. H. Garcia-Molina, "A Concurrency Control Mechanism for Distributed Databases Which Uses Centralized Locking Controllers"; submitted to the Fourth Berkeley Conference on Distributed Data Management and Computer Networks, August 1979.

7. H. Garcia-Molina, "Centralized Control Update Algorithms for Distributed Databases"; submitted to the First International Conference on Distributed Processing Systems, October 1979.

We will continue to track progress in this area, and when new results or systems appear assess their usefulness to a SUMEX-AIM like environment. We believe it remains important to look for more efficient ways to carry out the highly demanding computations in the area of artificial intelligence, and whereas improved heuristics and algorithms play a primary role in that search the exploitation of new hardware will complement such efforts.

4.2.5 MOLGEN Project

MOLGEN - An Experiment Planning System for Molecular Genetics

Profs. E. Feigenbaum and D. Lenat
Department of Computer Science
Stanford University

I. SUMMARY OF RESEARCH PROGRAM

A. Technical Goals

The MOLGEN project has constructed--and is extending--a computer system capable of generating the experiment-planning sequences needed to solve given structural problems in molecular genetics. In particular we have developed a system which is capable of acquiring and representing information about genetic objects, transformations, and strategies. The knowledge base presently includes information on DNA structures, restriction enzymes, laboratory techniques, and a growing collection of genetic strategies for discovering information about various aspects of DNA molecules. Several specific subproblems such as simulating Ligase enzymes, determining safe restriction enzymes for gene excision, and inferring DNA structures from segmentation data have been explored. We have designed our effort to facilitate generalization to other domains beyond genetics in future research and applications.

The MOLGEN project has both an applications and a computer science dimension. Along the latter dimension, we seek to deepen our knowledge of the art and science of creating programs that reason with symbolic knowledge at several levels (in this case, biological, genetic, topological, and chemical) to aid human problem solvers. To facilitate this, we have developed a knowledge representation system with a knowledge acquisition package. The system, known as the Units Package, may be used to build a knowledge base in any suitable domain. It provides an object-centered approach for storage of both declarative and procedural information concerning all entities in the domain. The task domain, molecular genetics, serves as a rich intellectual and scientific environment in which to develop and test our ideas.

The major computer science issues we are addressing during the current grant period are:

(1) Extending the UNITS package to efficiently represent a larger variety of knowledge; including four distinct notations for dealing with instances, schemata, descriptions, and variables; more meta-level information, nodes with multiple parents; explicit facts about the representation itself; and more efficient updating strategies for maintaining a very large knowledge base of units on disk.

(2) Creation of program schemata and instances for general problem solving steps. Domain-independent knowledge about general problem solving methods also fits into the knowledge representation structure we have devised.

(3) Domain Specific Critics. Mechanisms for the activation of various domain specific strategies when certain predefined situations occur during the course of experiment design.

(4) Development of a specific planning strategy designed to provide high-performance for the class of genetic experiments known as discrimination experiments. The idea is based on indexing abstracted experimental designs to the types of structural features for which they have proven useful.

(5) Constraint posting and Orthogonal planning. It has been observed that efficient planning proceeds hierarchically -- first making general decisions and later tending to the details. This research considers some planning problems with open-ended goals; part of the task of planning is to make the goals specific. A technique called "constraint posting" is presented for guiding the process of hierarchical planning by utilizing the interactions between parts of a plan. With this technique, a planning program can add details to a plan after recognizing and analyzing situations where more details are needed. The planner can skip around between different parts of the plan, suspending work on one section and moving to another rather than making uninformed decisions. The planner can accumulate constraints, which tie together decisions in separate parts of the plan. Other ideas include (1) a layered control structure termed "orthogonal spaces" which organizes the "plan-making" operations into explicit spaces separate from the domain knowledge and (2) some representations for hierarchical constants and variables which facilitate reasoning about abstract and undetermined objects during planning.

Along the applications dimension, we are attempting to develop tools that can benefit molecular geneticists. We believe there is substantial benefit to be derived from programs that act as "intelligent assistants" to scientists. First of all, the sheer amount of detailed knowledge a scientist is expected to know makes it likely that good experiments are being missed. Second, we believe that an intelligent planning assistant can offer some help in reasoning about the consequences of combining experimental facts in many possible ways.

A third motivation for applying artificial intelligence techniques to an experimental science like molecular genetics is to help us better understand the scientific method. The rigorous detail required for creating computer programs that assist in the performance of scientific tasks forces us to explicate concepts and procedures much more carefully than practicing scientists usually do.

B. Medical relevance and collaboration

Molecular genetics has at least two major connections to medical research. Learning about the basic mechanisms which control the operation and transmission of genetic information is necessary to understand and treat the wide range of diseases and health conditions that are genetically controlled. Also, recent developments in molecular genetics offer the promise of using genetic mechanisms to produce essentially limitless amounts of drugs and other biomedical substances.

The MOLGEN project is a joint effort of the Computer Science Departments of Stanford and the University of New Mexico and the Genetics Department of Stanford. Major participants are Professor Nancy Martin and Gary Klimowicz of the University of New Mexico; Professor Edward Feigenbaum, Professor Douglas Lenat, Professor Bruce Buchanan, Peter Friedland, and Mark Stefik of the Stanford Computer Science Department, Jerry Feitelson of the Stanford Genetics Department, Professor Laurence Kedes of the Stanford Medical School. Assistant Professor Douglas Brutlag of Stanford's Biochemical Department, Professor Joshua Lederberg, president of Rockefeller University and consulting professor of Computer Science at Stanford, and Dr. John Sninsky, a molecular biologist working in Professor Stanley Cohen's laboratory, are also collaborating in the MOLGEN project.

C. Progress summary

The major effort in MOLGEN has been the creation of a knowledge management system. In addition, several specific problems which arise in genetics have been examined in sufficient detail to result in reports and/or special purpose programs. We report briefly on two such programs, SAFE written by Peter Friedland, and GA-1 written by Mark Stefik.

Knowledge management system

The success of MOLGEN as an experiment planner will depend on the quality of its knowledge base. Therefore, much of the research effort to date has been in the design and implementation of a knowledge representation and acquisition system. All of the information relevant to the planning process will be an explicit part of the knowledge base. The motivation for this aspect of the design is the necessity to expand the program capabilities in a modular fashion and to explain the rationale behind the program's planning behavior. We need to represent concepts (e.g. enzyme), instances (e.g. EcoRI), relationships among concepts, and relationships among instances. In addition, we need to represent processes. We have purposely limited the expressive power of our representations to enable us to clearly define their semantics.

The result of this work is the Unit Package. Although this package has been designed in the context of our genetics application, the package does not contain any genetics knowledge.

One important aspect of the design of the system is that the knowledge base contains knowledge about its own data representations. We have provided what we term a "bootstrap knowledge base." It contains domain independent knowledge about commonly used data types. When using our knowledge base in a new domain, an artificial intelligence researcher would probably start with the bootstrap knowledge base and then proceed to create units for the specific knowledge of his task area. Both the AGE and genetics knowledge bases have been started in this manner. The bootstrap knowledge base serves to illustrate our approach to extensibility. Most of the bootstrap knowledge base is made up of primitive datatypes. To add a new datatype to our system, one needs to provide the knowledge base with procedures for some basic operations -- such as editing and printing. Actually, the same approach is used in the unit package for defining a new datatype as is used for defining a new enzyme. The process of defining new datatypes requires, however, an understanding of Interlisp because the primitive processes in the system are grounded in that language. New datatypes must be defined together with their basic operations and entered into the knowledge base.

Knowledge base contents

The genetics knowledge base is growing rapidly. Approximately 60% of the commonly used enzymes have been characterized. A beginning has been made on the characterization of organisms such as bacteria and phages, plasmids and other vectors, and genes. Our knowledge base also contains a growing collection of genetic strategies for discovering information about various aspects of DNA molecules, as well as a hierarchy of laboratory techniques which are used to instantiate the strategies. The hierarchy of techniques includes modification, separation, visualization, sequence analysis, and bacteriological techniques at many levels of abstraction.

Safe program

The geneticist needs to predict what restriction enzymes can safely be used to excise a gene, i.e. which ones can be guaranteed not to cut the functional part of the gene. We would also like to know the approximate location of the possible cutting sites of other restriction enzymes. This would all be very easy if the complete DNA sequence of the gene was known. Sequence information is becoming more and more prevalent, but it is still uncommon to know the complete sequence of a gene. However, it is not unusual to know what protein the gene codes for and to know the amino acid sequence of a protein.

Knowing the amino acid sequence does not provide full information because of the degeneracy in the genetic code. One codon (a triplet of nucleotides) specifies only one amino acid, but up to six different codons may specify the same amino acid. The problem therefore, is combinatorically difficult. Typical proteins are up to 300 amino acids long (900 nucleotides), and all possible nucleotide sequences which would produce the protein in the three possible phases have to be considered.

The SAFE program lists the restriction enzymes that are currently stored in the knowledge base and allows the user to add new ones. Besides determining which enzymes are safe to use for gene excision in a particular DNA molecule, the program also gives the position in the amino acid sequence where the possible cutting site would be located.

GA-1 program

A common task in molecular genetics laboratories is the analysis of DNA structure from restriction enzyme segmentation data. This task is one of the simplest, although time-consuming, analysis tasks in molecular genetics. Two standard approaches to solving this problem were examined: a data-driven strategy and a model-driven strategy. These approaches are discussed and compared in terms of sensitivity to missing data, efficiency in the use of data, and other measures of performance in [Stefik, 78]. A program was designed and implemented which is superior to human performance on smaller problems, both in speed and reliability. However, on large problems human problem solvers can use extra structural constraints to out perform the program. The current program uses only constraints derived from segmentation data itself. Geneticists usually know additional information -- eg. that a given segment is on the end or that certain segments must be adjacent.

A further benefit of this work was the suggestion of two new lab techniques: combining multiple enzyme digests and incomplete digests. These ideas arose from a systematic examination of evidence and inference rules that went into building the program.

D. Publications

Challenger J., A Program for Printing DNA Structures, CIS Report 78-3 (April 1978)

Feitelson J., Stefik M.J., A Case Study of the Reasoning in a Genetics Experiment, Heuristic Programming Project Report HPP-77-18 (Working Paper) (May 1977)

Friedland P., Knowledge-Based Experiment Design in Molecular Genetics, submitted to Sixth International Joint Conference on Artificial Intelligence (August 1979)

Martin N., Friedland P., King J.; Stefik M.J., Knowledge Base Management for Experiment Planning in Molecular Genetics, Fifth International Joint Conference on Artificial Intelligence. 882-887 (August 1977)

Stefik M., Friedland P., Machine Inference for Molecular Genetics: Methods and Applications, Proceedings of the National Computer Conference, (June 1978)

Stefik M.J., Martin N., A Review of Knowledge Based Problem Solving As a Basis for a Genetics Experiment Designing System, Stanford Computer Science Department Report STAN-CS-77-596. (March 1977)

Stefik M., Inferring DNA Structures From Segmentation Data: A Case Study, Heuristic Programming Project Report HPP-78-3 (January 1978) Stefik, M., An Examination of a Frame-structured Representation System, HPP-78-13 (Working Paper) (August 1978)

II. INTERACTIONS WITH THE SUMEX-AIM RESOURCE

All system development has taken place on the SUMEX-AIM facility. The facility has not only provided excellent support for our programming efforts but has served as a major communication link among members of the project. Through the SUMEX-AIM facility, program development has taken place concurrently at Stanford and New Mexico. Systems available on SUMEX-AIM such as INTERLISP, TV-EDIT, and BULLETIN BOARD have made possible the project's programming, documentation and communication efforts. The interactive environment of the facility is especially important in this type of project development.

We have taken advantage of the collective expertise on medically-oriented knowledge-based systems of the other SUMEX-AIM projects. In addition to especially close ties with other projects at Stanford, we have greatly benefitted by interaction with other projects at yearly meetings and through exchange of working papers and ideas over the system.

The combination of the excellent computing facilities and the instant communication with a large number of experts in this field has been a determining factor in the success of the MOLGEN project.

III. RESEARCH PLANS

A. Project goals and plans

In exploring the three major motivations mentioned in section I.A. for creating the MOLGEN project, there are many specific subproblems. We have identified five for concentrated effort in the next two year period.

(1) Creating a more comprehensive genetics knowledge base. Expanding the knowledge base within the area of DNA structural manipulation problems.

(2) Making use of the process of hypothesis formation to help debug MOLGEN-produced experiment designs. This process is especially important in a domain like molecular genetics where incomplete knowledge about objects and processes is the rule rather than the exception.

(3) Experiment planning by analogy. MOLGEN provides an excellent environment for exploring various types of analogical reasoning. We integrate problem-solving by analogy into the experiment design system as one of the possible tools for solving subproblems.

Each time the Heuristic Programming Project at Stanford has built another large AI program, we have learned more about how to do it better and faster next time. For example, the production rule interpreter in Heuristic Dendral (for special-purpose rules) became the general rule interpreter of MYCIN. One of the significant products of MOLGEN research will be the sets of ideas and programs for encoding and manipulating large amounts of knowledge about a scientific discipline. We have transferred some parts of the MOLGEN Units package to another project interested in building a knowledge base about AI methods and techniques. Making the tools used here available for use in new programs is an important aspect of our work and is generally important for cumulation of knowledge in the AI field. In order to do this we must reformulate the methods so they are more generally applicable and more readily combined in diverse ways.

B. Justification and requirements for continued SUMEX use.

The MOLGEN project is dependent on the SUMEX facility. While we have solved many of the original problems facing us in a manner useful to working geneticists, we are just in the middle phase of building a planning system. Without support from SUMEX to complete this system, many of the results of the last two years will be ineffective. In the past six months our interactions with geneticists outside of Professor Lederberg's laboratory have increased greatly. The geneticists are excited about helping us with our knowledge base. Also, with our help, they are finding useful ways to use the computing facility in their current research. Thus the serendipity of supporting MOLGEN is the creation of many useful computer research programs.

We are asked to state our requirements for continued SUMEX use. We project that our usage of processor cycles and file storage will grow to twice the current levels in the coming year.

4.2.6 MYCIN/EMYCIN Project

MYCIN/EMYCIN Project

Prof. B. G. Buchanan
Department of Computer Science
Stanford University

E. H. Shortliffe, M.D., Ph.D.
Department of Medicine
Stanford University Medical School

I. Summary of Research

A. Technical goals

MYCIN is an interactive consultation program which gives physicians antimicrobial therapy recommendations for patients with infectious diseases. The system must often decide whether and how to treat a patient before definitive laboratory results are available. It must recommend a therapeutic regimen which minimizes the risk of toxic side-effects while covering for all organisms which are likely to be causing the infection. The system currently has knowledge about treating three major infections - bacteremias (blood infections), cystitis and meningitis. The primary goal of the project has been to develop a program which would provide the same quality advice that a physician would get from a human infectious disease consult. Formal evaluations of the program's recommendations for patients with bacteremia and with meningitis have shown that this goal has been achieved.

Another important goal has been that the system should be easy to use and acceptable to a physician. To accomplish this, numerous human engineering features have been incorporated into the consultation, and an extensive explanation facility has been developed which enables the system to explain its reasoning and to justify its recommendations.

MYCIN's knowledge about infectious diseases is represented in production rules which are invoked by goal-directed backward chaining starting from the top-level goal of determining the appropriate therapeutic regimen.

The success of the MYCIN program in the area of infectious diseases led researchers to try to generalize and expand the methods employed in that program to a number of ends:

to develop consultation systems for other domains,

to explore other uses of the knowledge base,

to further facilitate interaction both for the developer of a knowledge-based system, and for the user of such a system, and

to experiment with using other knowledge representations in conjunction with the production rules used in MYCIN.

B. Medical Relevance

The MYCIN program was designed to help alleviate the problem of antimicrobial misuse documented in Shortliffe, 1976. We felt that MYCIN would be clinically useful when it was able to handle all major infections that are likely to be encountered in a hospital. It has not been possible to produce a clinically useful program in a few years due to the very large investment in time and human resources which are required of to develop, test and formally evaluate a rule set for each major infection area.

Through our collaboration on building the PUFF program, however, we learned that it is possible in a short period of time to develop a clinically useful consultation system using the domain-independent parts of the MYCIN program. Our effort to extract the "Essential" parts of MYCIN has been named EMYCIN; the development of EMYCIN can facilitate the creation of rule-based consultation systems for dealing with a number of medical problems.

C. Progress Summary

EMYCIN

Much of the work in the past year has been devoted to improving EMYCIN's facilities for allowing a system builder to construct and debug a knowledge base for a consultation system. This has included extensive documentation of the concepts used in EMYCIN consultation systems, the support programs for developing the knowledge base, and features of a working consultation system.

A knowledge-base debugging package was developed to assist the system builder in the task of testing, refining, and validating the knowledge base. This package includes: 1) the EMYCIN explanation facility; 2) a program that automatically explains how the system arrived at the results of a consultation; 3) a program that reviews each result of a consultation, allowing the user to judge whether the result is correct, and assisting the user in refining the knowledge base in order to correct any errors noted in the result or in intermediate conclusions; and 4) a program that automatically compares the results of a consultation to stored "correct" results for the same case, and explains any errors in the conclusions.

SACON

A new EMYCIN application called SACON was developed; this provided the EMYCIN developers with numerous ideas for improving EMYCIN facilities. A SACON consultation is meant to provide advice to a structural engineer regarding the use of a structural analysis program called MARC. The MARC program uses finite-element analysis techniques to simulate the mechanical behavior of objects. The engineer typically knows what he wants the MARC program to do--e.g., examine the behavior of a specific structure under expected loading conditions--but does not know how the simulation program should be set up to do it. The MARC program offers a large (and, to the novice, bewildering) choice of analysis methods, material properties, and geometries that may be used to model the structure of interest. From these options the user must learn to select an appropriate subset that will simulate the correct physical behavior, preserve the desired accuracy, and minimize the (typically large) computational cost. A year of experience with

the program is the typical time required to learn how to use all of MARC's options proficiently. The goal of the automated consultant is to bridge the "What-to-How" gap, by recommending an analysis strategy. This advice can then be used to direct the MARC user in the choice of specific input data--e.g., numerical methods and material properties. Typical structures that can be analyzed by both SACON and MARC include aircraft wings, reactor pressure vessels, rocket motor casings, bridges, and buildings.

The development of SACON represents a major test of the domain-independence of the EMYCIN system. Previous applications using EMYCIN have been primarily medical with the consultations focusing on the diagnosis and prescription of therapy for a patient. Structural analysis, with its emphasis on structures and loadings, allowed us to detect the small number of places where this medical bias had unduly influenced the system design, notably text strings used for prompting and giving advice.

The expert who provided the knowledge for the SACON program found that his knowledge was easily cast into the rule-based formalism and that the existing predicate functions and context-tree mechanism provided sufficient expressive power to capture the task of recommending an analysis strategy. The existing interactive facilities for performing explanation, question-answering, and consultation were found to be well developed and were used directly by our application. None of these features required any significant reprogramming and, for the most part, worked without modification.

GUIDON

Recent work has included the development of a tutoring system which uses the knowledge base of an EMYCIN consultation system as a manual of principles about the domain which can be taught to students. The tutorial program is domain-independent and can present whatever material is represented in the manual. The manual can be interpreted in two ways: (1) as a "runnable model" capable of performing the task to be learned by the student, and (2) as a set of principles to be discussed with the student.

The first version of GUIDON is being developed using the MYCIN manual of rules for diagnosing infections and prescribing antibiotic therapy. Other MYCIN-like manuals that are available include PUFF and SACON.

An important distinction between GUIDON and traditional computer-aided instructional programs is the independence of the tutor from the domain knowledge. As long as a manual is represented as a set of conditional rules (with the relevant objects identified and explicitly related), the tutor will be able to present the material. One area of investigation is the teaching of explicit problem-solving strategies that constitute approaches for applying the conditional rules to particular problems. Thus, this project represents a significant twist in knowledge-based AI research: we are taking knowledge that has been formalized in AI programs (the "manual") and TRANSFERRING IT BACK to humans, to students who want to learn the methods and strategies used by the experts in their field.

In separating teaching strategies from problem-solving strategies, we have explicitly stated the instructional methods we wish to test. Tutorial dialogue

knowledge is represented as procedures built from sequences of conditional rules. Thus, the teaching strategies for planning and directing the mixed-initiative dialogue can be readily displayed and changed. This will enable experimentation with alternate strategies, as well as making it easy to show them to other researchers.

Finally, GUIDON's mixed-initiative dialogue capabilities are more complex than in previous "intelligent computer aided instruction." The tutor engages the student in a dialogue while the student is working on a specific, complex task. A model of the student's knowledge in terms of the manual of rules guides teaching strategies for quizzing the student and presenting new information. The combination of a flexible environment for solving the problem (provided by the options we give the student for gaining more information and for keeping track of what remains to be done) and an active tutor that endeavors to convey problem-solving expertise, makes this a unique tutorial system.

D. Publications

Shortliffe, E.H., Axline, S.G., Buchanan, B.G., Merigan, T.C., and Cohen, S.N. An artificial intelligence program to advise physicians regarding antimicrobial therapy. Comput. Biomed. Res. 6,544-560 (1973).

Shortliffe, E.H., Axline, S.G., Buchanan, B.G., and Cohen, S.N. Design considerations for a program to provide consultations in clinical therapeutics. Proceedings of the 13th San Diego Biomedical Symposium, pp. 311-319, San Diego CA, 4-6 February 1974.

Shortliffe, E.H. MYCIN: A Rule-Based Computer Program To Advise Physicians Regarding Antimicrobial Therapy Selection. Doctoral dissertation, Stanford University, October 1974. Available as Technical Report HPP-74-2, Heuristic Programming Project, Stanford CA, October 1974. Abstract reproduced in Computing Reviews 16,385 (1975).

Shortliffe, E.H. MYCIN: A rule-based computer program for advising physicians regarding antimicrobial therapy selection. Proceedings of the ACM National Congress (SIGBIO Session), p. 739, November 1974. Reproduced in Computing Reviews 16,331 (1975).

Shortliffe, E.H., Rhame, F.S., Axline, S.G., Cohen, S.N., Buchanan, B.G., Davis, R., Scott, A.C., Chavez-Pardo, R., and van Melle, W.J. MYCIN: A computer program providing antimicrobial therapy recommendations. Presented at the 28th Annual Meeting, Western Society For Clinical Research, Carmel CA, 6 February 1975. Clin. Res. 23,107a (1975). Reproduced in Clinical Medicine, p. 34, August 1975.

Shortliffe, E.H. and Buchanan, B.G. A model of inexact reasoning in medicine. Math. Biosci. 23,351-379 (1975).

Shortliffe, E.H., Davis, R., Axline, S.G., Buchanan, B.G., Green, C.C., and Cohen, S.N. Computer-based consultations in clinical therapeutics: explanation and rule-acquisition capabilities of the MYCIN system. Comput. Biomed. Res. 8,303-320 (1975).

Shortliffe, E.H. Judgmental knowledge as a basis for computer-assisted clinical decision making. Proceedings of the 1975 International Conference On Cybernetics And Society, pp. 256-257, San Francisco CA, September 1975.

Davis, R. and King, J. An overview of production systems. Machine Intelligence 8: Machine Representations of Knowledge (eds. E.W. Elcock and D. Michie), John Wylie (1977). Also available as Technical Report HPP-75-7, Heuristic Programming Project, Stanford CA, October 1975.

Shortliffe, E.H. and Davis, R. Some considerations for the implementation of knowledge-based expert systems. SIGART Newsletter, No. 55, pp. 9-12 (1975).

Shortliffe, E.H., Axline, S.G., Buchanan, B.G., Davis, R., and Cohen, S.N. A computer-based approach to the promotion of rational clinical use of antimicrobials. Clinical Pharmacy and Clinical Pharmacology (eds. W.A. Gouveia, G. Tognoni, and E. van der Kleijn), pp. 259-274, Elsevier/North Holland Biomedical Press, New York, 1976.

Shortliffe, E.H. Computer-Based Medical Consultations: MYCIN, Elsevier/North Holland, New York, 1976.

Wraith. S.M., Aikins, J.S., Buchanan, B.G., Clancey, W.J., Davis, R., Fagan, L.M., Hannigan, J.F., Scott, A.C., Shortliffe, E.H., van Melle, W.J., Yu, V.L., Axline, S.G., and Cohen, S.C. Computerized consultation system for selection of antimicrobial therapy. Amer. J. Hosp. Pharm. 33:1304-1308 (1976).

Davis, R., Buchanan, B.G., and Shortliffe, E.H. Production rules as a representation for a knowledge-based consultation system. Artificial Intelligence, 8,15-45 (1977). Also available as Technical Report HPP-75-6, Heuristic Programming Project, Stanford CA, October 1975.

Davis, R. Applications of Meta-Level Knowledge to the Construction, Maintenance, and Use of Large Knowledge Bases. Doctoral dissertation, Stanford University, June 1976. Available as Technical Report HPP-76-7, Heuristic Programming Project, Stanford CA, July 1976.

Shortliffe, E.H. The MYCIN project: a computer-based consultation system in clinical therapeutics. Computer Networking In The University: Success And Potential, Proceedings of the EDUCOM Fall Conference, pp.183-185, November 1976.

Scott, A.C., Clancey, W., Davis, R., and Shortliffe, E.H. Explanation capabilities of knowledge-based production systems. American Journal of Computational Linguistics, Microfiche 62, 1977. Also available as Technical Report HPP-77-1, Heuristic Programming Project, Stanford CA, February 1977.

Buchanan, B.G., Davis, R., Yu, V., and Cohen, S.N., Rule based medical decision making by computer, Proceedings of MEDINFO 77, 1977.

Davis, R., Knowledge acquisition in rule-based systems: knowledge about representations as a basis for system construction and maintenance. Proceedings of Conference on Pattern-directed Inference Systems, May 1977.

Davis, R., Meta rules: content directed invocation. Proceedings of ACM conference on AI and Programming Languages, August 1977.

Davis, R., Interactive transfer of expertise: acquisition of new inference rules. Proceedings of Fifth IJCAI, August 1977.

Davis, R., and Buchanan B.G., Meta level knowledge: overview and applications. Proceedings of Fifth IJCAI, August 1977.

Aikins, J. The Use of Models in a Rule-Based Consultation System. Proceedings of Fifth IJCAI, P. 788, August 1977.

Shortliffe, E.H. A rule-based approach to the generation of advice and explanations in clinical medicine. In "Computational Linguistics in Medicine" (eds. W. Schneider and A.L. Sagvall Hein), pp 101-108, North Holland, Amsterdam (1977).

Shortliffe, E.H. MYCIN: A knowledge-based computer program applied to infectious diseases. Proceedings of the Symposium on Computers in Medical Care Delivery, pp 66-69, Washington D.C., October 1977.

Shortliffe, E.H. Clinical decisions based on physician-computer interactions: a symbolic reasoning approach. Symposium on Making and Using Medical Decisions, Proceedings of the annual meeting, Society for Computer Medicine, Las Vegas, Nevada, November 1977.

Davis, R., A decision support system for medical diagnosis and therapy selection, in Data Base (SIGBDP Newsletter), 8:58 (Winter 1977).

van Melle, W. MYCIN: A knowledge-based consultation program for infectious disease diagnosis. Int. J. Man-Machine Studies 10, 313-322 (1978).

Bonnet, A. BAOBAB, A parser for a rule-based system using a semantic grammar. Technical Report HPP-78-10, Heuristic Programming Project, Stanford CA, September 1978.

Bennett, J.S., et al. SACON: A Knowledge-based Consultant for Structural Analysis, Department of Computer Science, Stanford University, Memo HPP-78-23, Sept. 1978.

Clancey, W. The Structure of a Case Method Dialogue, International Journal of Man-Machine Studies, Fall 1978.

Shortliffe, E.H. Interactive programs for physicians: benefits and limitations of artificial intelligence techniques. Proceedings of the International Conference on Cybernetics and Society, p. 302, Tokyo, Japan, 3-5 November, 1978.

Shortliffe, E.H., Buchanan, B.G., and Feigenbaum, E.A. Knowledge engineering for medical decision making: a review of computer-based clinical decision aids. Submitted for publication in the Proceedings of the IEEE, December 1978.

Yu, V.L., Buchanan, B.G., Shortliffe, E.H., Wraith, S.M., Davis, R., Scott, A.C., Cohen, S.N. Evaluating the performance of a computer-based consultant. Computer Programs in Biomedicine, 9,95-102, (1979).

Yu, V.L., Fagan, L.M., Wraith, S.M., Clancey, W.J., Scott, A.C., Hannigan, J.F., Blum, R.L., Buchanan, B.G., Cohen, S.N. Antimicrobial selection for meningitis by a computerized consultant -- a blinded evaluation by infectious disease experts. To appear in JAMA, 1979.

van Melle W., A domain-independent production-rule system for consultation programs, submitted to 6IJCAI (1979).

II. Interaction With the SUMEX-AIM Resource

A great deal of interest in MYCIN has been shown by the medical and academic communities. Among the people who have visited the project or asked for GUEST access to the MYCIN program are Dr. Solveig Pflueger of the University of Texas at San Antonio Medical School, Dr. Robert H. Rosenberg of Seattle Radiologists, Inc., and Dr. Jeffrey P Krischer, Chief of Health Services Research and Development at the University of Florida. Dr. Peter Szolovitz of MIT and Dr. Steven Zucker of McGill University in Montreal have demonstrated the MYCIN program in their university classes. Dr. Harold Goldberger of MIT made extensive use of the MYCIN program last last summer in his study of medical AI programs. Dr. Ves Morinov of the Norwegian Computing Center has used the MYCIN program to demonstrate the benefits of using a rule-based representation for consultation systems. The MYCIN program was demonstrated at the Third Rheumatology - Information Science Meeting in March 1978, at the Annual meeting of the American College of Physicians in San Francisco in March 1979, and at a monthly meeting of the Stanford Computers in Medicine Group in May 1979.